a Ground the scene

Multi-view RGB







Main + left / right wrist cameras

"Place the mug in the tray."

Robot state

SigLIP encoder

PaliGemma

Vision–language backbone

Target · Destination · Obstacles

4 learned tokens per semantic role

Role-conditioned mask decoder



Target · Destination · Obstacles

b Generate grounded actions

Visual + language + state + mask role context

Early layers

Middle layers

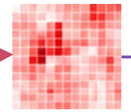
Late layers
Geometry fusion

Attention residual

Read attention

Action → image attention
Average heads + queries
AttnQueryHead

KL supervision (d)



Learned maps

Average, detach & inject
Mechanism (e)

f Execute



50-step arm + gripper action chunk

c Extract object depth and fuse geometry

Main-view RGB



VGGT-Ω

Scene depth + confidence



Predicted masks (from a)





Align, detach & threshold masks
Filter depth by confidence
Pool within each role

Selected object depth



Target



Destination

Per-role depth encoding

Image position
+ Z statistics

4 × 4 mean-Z grid

Learned projector + role embeddings

34 geometry tokens

17 target

17 destination

1 summary + 16 grid / role

Geometry tokens enter the action suffix; late layers attend to their valid keys. Predicted target and destination masks select the depth regions.

d Learn where to attend TRAINING ONLY

Expert geometry



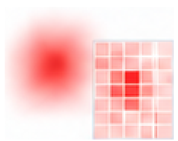
Trajectory + obstacle labels

Closest approach



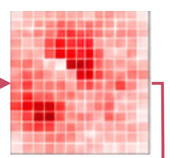
FK, contact anchors + β

Weighted contact mask



Gaussian blur + patch pooling

Target P_{GT}

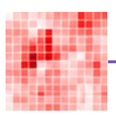


16 × 16 patch grid

$$\mathcal{L}_{attn} = \text{mean}_{\ell}\, w_{\ell}\, D_{KL}(P_{GT} \,\|\, Q_{\ell})$$

Contact labels and β supervise attention during training.

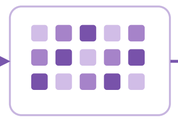
e Inject learned attention into late action layers



Q_1, Q_2, Q_3

Layer average
Stop gradient

3 × 3 conv + position
Linear projection



Spatial token bank

K, V

Action tokens h

Query

Cross-attention

Scale × Δh

h′

h′ = h + scaled attention update

Stop-gradient isolates the injection branch. KL supervision trains the source maps through (d).